Deep Learning on Indoor Localization: A Million-Scale Database with Millimeter-Level Labels

ACM MobiCom 2022, Paper#1007

ABSTRACT

High-precision tracking inside buildings where GPS is unavailable has become a game changer for various industries, such as virtual reality and factory automation. Unlike outdoor localization, which has thrived over the past decades because of the well-established global position system, indoor localization systems underperform the overall market because of many long-standing issues, such as the hardware and spatial diversity, multipath effect, and so on. We demonstrate that the indoor localization can be reduced to a joint non-linear optimization problem that can be conquered by a deep neural network (DNN). Unfortunately, the current existing localization databases cannot meet the quality and quantity requirement for deep learning. In this work, we use a customized acquisition platform to establish a million-scale database with millimeter-level labels, called *Ray* database, which can be employed as a general-purpose and protocol-free dataset for benchmark or network training. We also propose a novel DNN for indoor localization, called *ThreeBodyNet*, to demonstrate the potentials of the proposed database. ThreeBodyNet outperforms state-of-the-art methods in indoor localization across 17 different scenes.

1 INTRODUCTION

Recent years have witnessed an upsurge of interest in developing accurate RF-based indoor localization systems to compensate for the last hundred meters where GPS service is unavailable. To this end, many efforts [1–19] have been made in recent two decades, which have achieved a rather encouraging accuracy. High-precision indoor localization can facilitate a series of key applications, such as indoor navigation, augmented reality, location-aware pervasive computing, advertising, social networking, and so on. Consequently, tracking IoT devices inside buildings have become a growing business interest.

Unfortunately, few previous works were really deployed in the real world due to various unexpected on-site negative factors across deployment scenes. (1) *Hardware diversity*: the diversity of hardware caused by the circuitry characteristics may introduce unwanted errors in measuring RF signals, shifting the input of a localization algorithm with many unknown offsets. (2) *Spatial diversity*: the RF signal varies across space due to the uneven distribution of electromagnetic fields. (3) *Multipath effect*: in the face of an uncertain, complex, and dynamic wireless environment, RF signals might be reflected from many static or moving obstacles. The multipath effect has become the greatest threat to indoor localization when the localization algorithm highly depends on the line-of-sight (LoS) propagation. (4) *Coordinate conversion*: a localization system is a distributed system that deploys many stations (i.e., gateways) in different positions. Each station can estimate a measurement (e.g., direction) on the device relative to its local coordinate system. These intermediate measurements must be converted into a result in the global coordinate system. Such conversion requires another high-precision system (e.g., optical position system) to pinpoint the stations first and then integrate them into a whole system. Particularly, the last two factors are highly dependent on the deployment scenes. They prevent the developed and tested prototypes in a lab from being widely deployed in a real-life scene.

As modeled in §2, the indoor localization can be reduced to jointly resolving two related non-linear optimization problems where the above factors are the certain variables or constants. This problem falls into the domain of machine learning. Thus, many previous works started to resort the advance from the deep neural network (DNN) to address indoor localization [21, 25–28], which have successfully demonstrated the outperformance of deep learning in accuracy and stability compared with conventional localization algorithms.

A growing consensus exists in the industry and academia that the performance of deep learning highly depends on the quantity and the quality of training samples. We summarize the existing database developed for the RF-based indoor localization. Specifically, DLoc [25] offers the greatest number of records of up to 105 k across two scenes; Microsoft database involves the most number of scenes (i.e., 212), but the label is room level and annotated manually; WiDar, Intel, and DLoc database employed LiDAR for the label acquisition, and thus their label precision is limited to cm level; the first four databases are collected in a single scene, and the area of the largest scene is limited 185 m^2. The previous existing databases suffered from many limitations in the data scale, the scene diversity, the label precision, and the coverage. They are far beyond as a general-purpose database for benchmark, training and transfer learning regardless of the quality or the quantity.

This work introduces Ray, a first million-scale 3D indoor localization dataset with millimeter-level labels. We customize a general acquisition platform equipped with three 4×4 antenna arrays. Although the data are collected from

Table 1: Summary of existing databases for RF based localization

Database	Year	Target	Frequency	Samples (#)	Area (m^2)	Scenes (#)	Label	Label Precision
Toshiba [20]	2019	BLE UWB	2.4 GHz	1.6 k	16.2	1	OptiTrack	0.1 mm
iArk [21]	2020	IoT	800-900 MHz	28 k	200	1	OptiTrack	0.1 mm
Widar [22]	2019	Wi-Fi	2.4 GHz	30 k	N.A.	1	LiDAR	10 cm
Intel [23]	2020	Wi-Fi	2.4 GHz	30 k	750	1	LiDAR	30 cm
Microsoft [24]	2021	BLE, Wi-Fi	2.4 GHz	100 k	N.A.	212	Manually	Room-level
DLoc [25]	2020	Wi-Fi	2.4 GHz	105 k	185	2	LiDAR	5.7 cm
Ray	2022	RFID	800-900 MHz	1,323 k	7,854	37	OptiTrack	0.1 mm

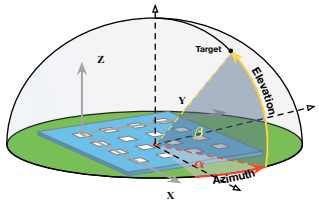


Fig. 1: Local Coordinate Systems

RFID tags, they can be employed as a protocol-free and general-purpose database for training any type of localization neural network, due to the dual-channel signal processing in the baseband. Sponsored by the top world-class communication company, we invested over 200 K USD and exerted two-year efforts on the hardware development, manpower, and data collection.

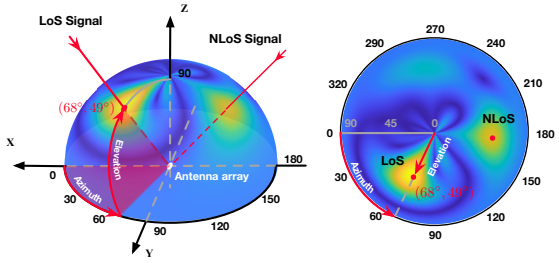
We briefly summarize the main merits of the Ray database as follows: **(1) Data Scale**. To the best of our knowledge, Ray is the largest localization database with over 1.32 million location records, i.e., 13 times larger than DLoc and Microsoft databases in scale. **(2) Coverage and Spatial Diversity**. We collected the Ray database from 14 different scenes with 37 different settings (such as various distances or layouts). The largest coverage is up to $7,854\ m^2$, which is 10 times larger than the past largest. **(3) Label Precision**. We adopt optical positioning system (e.g., OptiTrack) to acquire the labels at the millimeter precision. We specially develop a queue system to synchronize the labels and the packets acquired by the gateways. More property analysis is presented in §6.

In this work, we also propose a new DNN, *ThreeBodyNet*, which exhibits as an example to demonstrate the potentials of the Ray database in §7. The DNN is named after its particular network architecture, which comprises three body networks (subnetworks). Each body network is created via a ResNet-50 based convolutional neural network. Correspondingly, we propose four strategies to train ThreeBodyNet using our Ray database. Resultantly, ThreeBodyNet can achieve a median error of 12.2 cm and a 90th percentile error of 33.1 cm on average across all kinds of full-indoor environments, which outperforms by 56.4% than the state-of-the-art localization neural network.

Contribution. Our contributions are summarized as follows. First, we build an acquisition platform consisting of three distributed gateways, which collect RF signals at three positions; second, we establish an ultra large-sized database for localization using the platform; third, we propose a DNN for indoor localization to achieve a satisfying accuracy.

2 PRELIMINARY

This section introduces the background knowledge about the spatial spectrum and the triangulation-based localization.



(a) 3D Spatial Spectrum (b) 2D Spatial Spectrum

Fig. 2: Illustration of spatial spectrum

2.1 Spatial Spectrum

Suppose the antenna array is equipped with $K \times K$ elements uniformly where two adjacent elements are spaced with L. As shown in Fig. 1, we can set up a local *cartesian coordinate system* for an array by choosing its bottom-left corner as the origin. Choosing the element $E_{1,1}$ as a reference, we can compute the following relative power of projecting the received signal into the direction of (α, β):

$$P(\alpha,\beta) = \frac{1}{(K^2-1)} \left| \sum_{i=1}^{K} \sum_{j=1}^{K} w_{i,j}(\alpha,\beta) \cdot e^{\mathrm{J}\Delta\widetilde{\theta}_{\mathrm{i,j}}} \right| \tag{1}$$

where $w_{i,j}(\alpha,\beta) = e^{\mathrm{J}-\Delta\theta_{\mathrm{i,j}}}$ is the complex weight for steering a beam to a certain angle of (α,β). In the above, $\Delta\widetilde{\theta}_{i,j}$ is the true phase difference computed by using the received signals at $E_{i,j}$ and $E_{1,1}$, whereas $\Delta\theta$ is their theoretical phase difference. The sum aggregates the relative power across the (K^2-1) pairs of elements, i.e., $(E_{1,2}, E_{1,1})$, $(E_{1,3}, E_{1,1}), \cdots$. When $\Delta\widetilde{\theta}_{i,j}$ aligns with $\Delta\theta_{i,j}$ (i.e., the signal does come from the direction of (α,β)), the normalized relative power $P(\alpha,\beta)$ should achieve the maximum. For clarity, we use ω to denote the tuple of the two angles related to a direction, i.e., $\omega = (\alpha,\beta)$. The relative power at the direction ω is rewritten in the form of the vector as follows:

$$P(\omega) = \left[w_{1,1}(\omega), w_{1,2}(\omega), \cdots, w_{K,K}(\omega) \right] \left[e^{\mathrm{J}\Delta\widetilde{\theta}_{1,1}}, e^{\mathrm{J}\Delta\widetilde{\theta}_{1,2}}, \cdots, e^{\mathrm{J}\Delta\widetilde{\theta}_{\mathrm{K,K}}} \right]^T$$

where T denotes the transpose.

Next, a hotmap can be generated to show the relative power at N possible directions that the received RF signal might come from. We call such a hotmap *spatial spectrum*, which is formalized as follows:

$$\Omega = \mathbf{W}\mathbf{S} + \mathbf{Z} \tag{2}$$

where Ω, $\mathbf{W}$, $\mathbf{S}$, and $\mathbf{Z}$ denote the spatial spectrum, the weight matrix, the received signals, and the noise, respectively. N is a custom parameter depending on the angle resolution. $N = 360 \times 90$ if one degree resolution is accepted. Fig. 2 shows an example of the spatial spectrum. Particularly, Fig. 2(a) shows the spatial spectrum in 3D where all directions are uniformly distributed; Fig. 2(a) shows the 2D spectrum by projecting the 3D onto the X-Y plane, in which the radial distance represents $\cos(\beta)$ so the elevation angle distributes non-uniformly. The spatial spectrum should peak at the direction that the RF signal truly comes from. Formally, the direction of the device is computed by solving the optimization problem as below:

$$\omega* = \underset{\omega}{\text{argmax}}\ \Omega = \underset{\omega}{\text{argmax}}\ (\mathbf{WS} + \mathbf{Z}) \tag{3}$$

It is worth noting that the community proposed many different types of spatial spectrums by using different weight matrices, such as Bartlett [29], MVDR [30], MUSIC [31], and Tagoram [1]. There, the Bartlett is adopted and others are discussed in the evaluation.

2.2 Triangulation

At least two antenna arrays are deployed in the target space to compute the location of the device. Specifically, we can compute a direction using a single antenna array. The device is located at the intersection of two directions or the centroid of the intersected area formed by multiple directions. This localization approach is called *triangulation*. As aforementioned, the direction of the device is estimated relative to the local coordinate system (LCS) of an array. Thus, it requires an extra step to convert all directions computed in local coordinate systems to a global coordinate system (GCS) before taking the triangulation. Formally, suppose the estimated direction $\omega^* = (\alpha^*, \beta^*)$ by using an antenna array, a straight line denoted by l in the LCS relative to the array can be constructed as follows:

$$l(\vec{a}(\omega^*), O) \tag{4}$$

where $\vec{a}$ and O are the directional vector of the line and the origin of the LCS that it passes through, namely,

$$\vec{a}(\omega^*) = [1, \tan(\alpha^*), \tan(\beta^*)/\cos\alpha^*]^T \textbf{ and } O = [0, 0, 0]^T \tag{5}$$

To compute the intersection, we firstly convert the above line from the LCS to the GCS using a rotation matrix $\mathbf{R} \in \mathbb{R}^3$ and the coordinate of the array $\mathbf{O}$ in GCS. Then, the straight line in the GCS is rewritten as follows:

$$l(\mathbf{R} \cdot \vec{a}(\omega^*), \mathbf{O}) \tag{6}$$

Suppose the directions are estimated by two antenna arrays in different positions. Then the location of the device is to solve the following optimization problem:

$$p^* = \underset{p}{\text{argmin}} \sum_{g=1}^{G} \mathbf{d}(p, l(\mathbf{R}_g \cdot \vec{a}(\omega_g^*), \mathbf{O_g})) \tag{7}$$

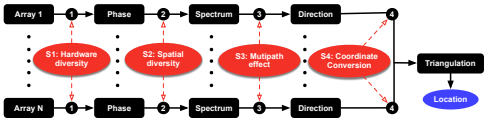


Fig. 3: Localization workflow and error sources

where $\mathbf{d}(\cdot)$ is the Euclidean distance between a point and a line, and G is the number of antenna arrays.

3 WHY DEEP LEARNING HELPS?

The goal of indoor localization can be reduced to resolving two optimization problems shown in Eqn. 3 and Eqn. 7, which exactly falls into the domain of machine learning. However, it becomes extremely complicated because of the unexpected all-pervasive error sources, making the problem far beyond the capacity of traditional optimization approachees.

3.1 Error Sources

The triangulation-based localization works well in the free space without obstacles or reflectors, especially when the devices are far away from the arrays. Thus, it has been adopted in the Bluetooth 5.1 [32] and mm-wave systems [33]. However, this approach suffers from the serious "butterfly effect" due to the geometric nature of triangulation. Namely, a small error occurred in any step is gradually magnified and finally leads to a tremendous error in a localization result, i.e., a miss is as good as a mile. Thus, we firstly sort out all sources that might trigger the butterfly effect in the triangulation. Fig. 3 illustrates the whole localization workflow, including four main steps. Specifically, ❶ estimating the phase values of the incoming RF signals at all elements of an antenna array (i.e., resolving S); ❷ aggregating the phase to generate a spatial spectrum (i.e., resolving Ω); ❸ computing the direction as the angles spike at the spatial spectrum (resolving $\omega*$); ❹ determining the location as the intersection of two or more directions (resolving $p*$). The intermediate errors might occur at any one of the above steps as follows:

Source 1: Hardware Diversity. We assume that the phase rotation is only caused by the distance. Actually, the phase values are also affected by the hardware diversity caused by the uneven wiring or the coupling effect [1]. Such diversity introduces a constant offset to the measured phase at each antenna. We use a diagonal matrix $\mathbf{C}$ to denote the offsets at each antenna. The spatial spectrum is rewritten to:

$$\Omega = \mathbf{WCS} + \mathbf{Z} \tag{8}$$

Traditionally, each array must be calibrated for once after it is fabricated such that the $\mathbf{C}$ can be compensated.

Source 2: Spatial Diversity. We also observe the spatial diversity in antennas. Ideally, the measured phase should remain unchanged if the distance between the RF source and the antenna is held. Actually, we find that the phase varies in a roughly 0.8 rad range as a function of the azimuth angle.

This spatial diversity is resulted from the uneven distribution of the electromagnetic fields. Let $\mathbf{\Gamma}$ be the calibration vector at different angles. The spatial spectrum is further rewritten to:

$$\Omega = \mathbf{WC\Gamma}(\omega)\mathbf{S} + \mathbf{Z} \tag{9}$$

Source 3: Multipath Effect. The most notorious error source for indoor localization is the multipath effect, namely, multiple copies of the RF signal caused by the reflections from the surrounding obstacles are also received by arrays. For clarity, we assume that the signals travel through M paths. $\mathbf{S^m}$ represents the signal propagated along the m^{th} to the $K \times K$ antennas. Consequently, the spatial spectrum is written as follows:

$$\Omega = \mathbf{WC}\left(\Gamma(\omega_0)S^0 + \sum_{m=1}^{M}\Gamma(\omega_m)S^m\right) + \mathbf{Z} \tag{10}$$

Note that the multi-path propagations are from different directions, so the spatial diversities are not identical. For simplicity, we assume that S^0 denotes the signal propagated from the line-of-sight path and there are M multiple paths. Clearly, only the S^0 is useful in the following triangulation and other propagations are noises. After obtaining the spatial spectrum from an antenna array, the device is determined as the direction at which the spectrum peak. Correspondingly, the optimization problem is updated as follows:

$$\omega^* = \underset{\omega}{\text{argmax}}\ \Omega = \underset{\omega}{\text{argmax}}\ \mathbf{WC}\left(\Gamma(\omega_0)S^0 + \sum_{m=1}^{M}\Gamma(\omega_m)S^m\right) + \mathbf{Z} \tag{11}$$

Source 4: Coordinate Conversion. Finally, after the system is deployed, another set of high-precision localization system (i.e., optical localization system) must be leveraged to set up the global coordinate system and measure the normal vectors and positions of the arrays. However, this procedure would introduce uncontrollable errors, such as operating error, interference, ambient error, and so on. In other words, the two variables of $(\mathbf{R}, \mathbf{O})$ in Eqn. 7 are inaccurate in practice.

3.2 Motivation behind Deep Learning

The discussion so far focuses on the potential error sources which might introduce errors leading to the butterfly effect. Regarding these errors, the ultimate objective function is now complicated to jointly resolve the two optimization problems defined in Eqn. 11 and Eqn. 7, which involves four *unknown* and *hidden* variables in the form of matrix or vector, i.e., $\mathbf{C}$, $\mathbf{\Gamma}$, $\mathbf{R}$, and $\mathbf{O}$. Notably, the two equations are hardly resolved out without plenty of measurements.

Particularly, **Source 1**, **Source 2**, and **Source 4** introduce *static error variables*, i.e., $\mathbf{C}, \mathbf{\Gamma}$, and $(\mathbf{R}, \mathbf{O})$. They are introduced by the hardware characteristics and the deployed positions in the scene. These constants can be "learned" by a DNN from a large number of training samples. **Source 3** introduces *dynamic variables*, especially when multiple reflectors exist around. Ideally, the line-of-sight (LoS) propagation can be separated from the received signal if we can sort out the pattern of signal propagation in a particular scene. This is technically sound for two main reasons. First, the multipath propagation (i.e., none line-of-sight, NLoS) travels a longer distance, and thus its signal is weaker than that of the LoS signal. Second, most of the NLoS signals are created by the major reflectors (such as the walls, ceilings, furniture, and so on), whose influence can also be modeled by a ray tracing algorithm ideally. The accurate ray tracing is impractical because it demands the accurate geometrical information about the reflectors. Nevertheless, the machine can still "learn" the implicit models of the environment if given a vast number of samples.

At the heart of a DNN is to solve a non-linear optimization problem like ours by using numerous training samples. A DNN, especially the convolution neural network (CNN), is known for discovering the inherent and hidden variables robust to the ambient interference, skews, and distortions in input data. Given such task comparability, we believe that the deep learning can help resolve this long-standing issue in indoor localization. The previous work [21, 25–28] also show some potentials of deep learning in indoor localization.

4 OVERVIEW

Applying deep learning in indoor localization is promising. However, it is currently suffering from a severe shortage of well-documented and high-precision training datasets. To the best of our knowledge, there are few high-quality benchmark databases for this end. Thus, we develop a platform to collect the millions of locations with millimeter-level labels in dozens of scenarios after two-year efforts. A total of 200 K USD were spent on the acquisition platform, manpower, and data collection. The database will be publicly available and free to all industrial and academic researchers once this work is accepted (for keeping the anonymity). In the following, we first introduce the acquisition methodology, including designs of the RF platform and the label collection system in §5. We present the collected datasets and comprehensively analyze their properties and diversities in §6. Finally, we propose a new DNN to demonstrate the way of using our database to train a neural network in §7. Next, we will elaborate on these issues.

5 ACQUISITION METHODOLOGY

First, we introduce our hardware, a distributed localization platform, which is employed to acquire the signals from RFIDs at different positions. Although the signals are acquired from RFIDs, our platform is protocol-free and works for any devices operating at 800-900 MHz. The acquired data (regardless of the IQ data or phase values) are independent on the protocols and thereby can be used as a general-purpose

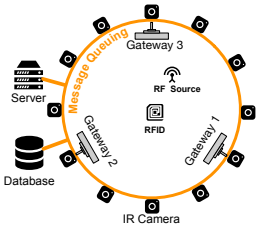


Fig. 4: Deployment

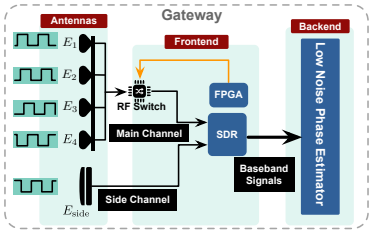


Fig. 5: Architecture of the gateway

dataset for training the neural network. Meanwhile, a set of submillimeter-accuracy optical positioning system is deployed to collect the ground truth.

5.1 Architecture

Fig. 4 shows the architecture of our acquisition platform. The whole platform comprises two or more gateways (equipped with antenna arrays), an optical positioning system (OPS), a sever and a database. All these components are connected through a message queuing (MQ) system. The surveillance region is surrounded by 3 gateways and 12 infrared cameras. **(1) Gateway**: the core components in the hardware layer are the gateways, each of which is equipped with a 4×4 antenna array. A gateway keeps listening for the RF signals as a sniffer. When receiving a packet successfully, the gateway estimates the values of (RSS, Phase) at each element and then delivers the 16 pairs of (RSS, Phase) to the corresponding gateway queue. **(2) OPS**: We adopt a high-precision OPS (e.g., OptiTrack [34]) to obtain the ground truth (aka labels). The system requires to deploy multiple infrared (IR) cameras around the region, which can track an IR marker with an accuracy of $20\mu m$. An IR marker is attached to the center of the RFID tag. The OPS can report the ground truth at a rate of 120 Hz. **(3) Server and Database**: the server maintains a spectrum queue for each gateway and a label queue for the OPS. The queues are used to store the spatial spectrums generated from the gateways and the ground truth from the OPS. After the message alignment, all data are saved into the database. **(4) MQ System**: all above components are connected through a message queuing system (e.g., RabbitMQ [35]). **(5) Device of Interest (DOI)**: we deploy an RFID tag and an RF source on a robot, which randomly moves in the surveillance region. The RF source is to power up the target battery-free RFID tag and uses `Select` and `Query` commands to guide the tag to reply. Although we employ the RFID tags as the DOI to establish the database, we must claim that the collected data actually are free of any wireless protocol.

5.2 Acquiring RF Signals via Gateway

At the heart of a gateway is the dual-channel design, i.e., a main channel and a side channel. As shown in Fig. 5, the main channel equipped with an antenna array is to acquire

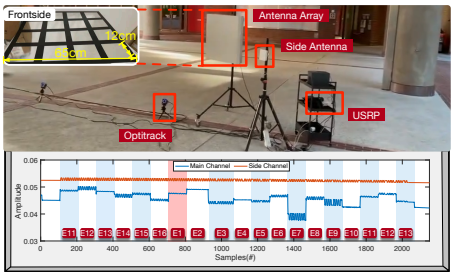


Fig. 6: Photo of the gateway and the baseband signal acquired by the gateway

the RF signal from different positions, while the side channel equipped with a side antenna is to decode the packet piggybacked by the signal.

• **Main Channel**: the main channel comprises a uniform planar antenna array with 4×4 elements and an ADC. Taking into account the high cost of high-speed ADC, we use RF switches to bridge one of the antenna elements to the ADC in a time-sharing fashion. The switching is controlled by an FPGA. Since the arrival time of a packet is unknown, the FPGA puts through an element at regular intervals and schedules the switching cyclically and persistently. However, at any time, only a single element is connected to the ADC. In this way, the packet transmitted from the target device can be captured by the 16 elements, each of which captures a segment of the packet only.

• **Side Channel**: the side channel comprised a side antenna and another ADC. It can persistently capture RF signals continuously. The backend leverages the signal captured by the side channel to detect the presence of a packet. If the packet is detected and decoded successfully, the gateway will trigger the phase estimation and spectrum generation.

Implementation. Fig. 6 shows the prototypes of the gateway. We use the microstrip technique to fabricate the antenna array on a printed circuit board (PCB). The PCB comprises a substrate of RT/duroid 5880. The model of high-speed RF switches is HMC241 [36]. The size of the antenna array is 65×65 cm^2, each of which is 12×12 cm^2. Each array element has 0.5 dBi gain, 0.5 dB flatness, and ±45° dB beamwidth. Fig. 6 also shows the baseband signals acquired concurrently by the two channels from an RFID tag's `RN16` reply. In the main channel, the received signal is sliced into segments through the amplitude changes. The E1 segment with a larger window is used to determine the beginning of the schedule. Before or after E1, other 15 equal segments are received by the antennas $E_2 \sim E_{16}$, respectively.

5.3 Baseband Processing

Let $X(t)$ and $C(t)$ denote the baseband signal and the carrier signal, respectively. After the modulation, the transmitted signal $S_T(t)$ propagated into the air can be modeled as:

$$S_T(t) = X(t)C(t) = X(t)e^{\mathrm{J}2\pi f_c t} \tag{12}$$

where $C(t) = e^{\mathrm{J}2\pi f_c t}$. At the receiver side, the received signal differs from the transmitted signal due to the propagation in the air. The received signal $S_R(t)$ turns into:

$$S_R(t) = HS_T(t) = aX(t)e^{\mathrm{J}(2\pi f_c t+\theta)} \tag{13}$$

where $H = ae^{\mathrm{J}\theta}$ is the channel parameter, a and θ are the amplitude attenuation and the phase rotation caused by the propagation in the air. Similar to the modulation, the receiver multiplies the received signal by the conjugate of a carrier signal $C_R(T) = e^{\mathrm{J}2\pi f_c' t}$ for demodulation. Then, the demodulated baseband signal $Y(t)$ is given by:

$$Y(t) = S_R(t)C_R^*(t) = aX(t)e^{\mathrm{J}(2\pi\Delta f_c t+\theta)} \tag{14}$$

where $\Delta f_c = f_c - f_c'$ is called the carrier frequency offset (CFO) because of the frequency agnosticism. Our purpose is to estimate the θ, which is highly related to the position of the target device. Unfortunately, due to the presence of CFO of baseband signal, the estimated phase value becomes a time-varying variable:

$$\widetilde{\theta} = 2\pi\Delta f t + \theta + \angle X(t) \tag{15}$$

Essentially, the CFO is a result of the out-of-sync in clock between the transmitter and the receiver. However, the signals received by our dual channels are affected by the same CFO since the signal is transmitted from a same device. Inspired by this insight, we take the side channel as a baseline to eliminate the CFO. Let $Y_0(t)$ and $Y_{i,j}(t)$ be the signals received by the side channel and the array element $E_{i,j}$ in the main channel. The relative phase is defined as follows:

$$\begin{aligned}\widetilde{\theta}_{i,j} &= \angle Z_{i,j}[t] = \angle\left(\frac{Y_{i,j}[t]}{Y_0[t]}\right) = \angle\left(\frac{a_{i,j}\cancel{X[t]}e^{\mathrm{J}(\cancel{2\pi\Delta f t}+\theta_{i,j})}}{a_0\cancel{X[t]}e^{\mathrm{J}(\cancel{2\pi\Delta f t}+\theta_0)}}\right)\\ &= \angle\left(\frac{a_{i,j}}{a_0}e^{\mathrm{J}(\theta_{i,j}-\theta_0)}\right) = \theta_{i,j} - \theta_0\end{aligned} \tag{16}$$

As the above equation indicates, both two uncertain variables are removed from the division. As a common constant, the θ_0 will also be removed when computing the relative power in Eqn. 1. Hundreds of phase values are estimated out of a single packet and the average is reported. Since an array contains 16 elements, the 16 phase values as an array is saved.

Implementation. We use a USRP X310 software-defined radio (SDR) from NI [37] for the baseband signal processing of a gateway. Specifically, each X310 is equipped with two TwinRx daughterboards. One RX channel is connected to the antenna array, while another RX channel is connected to a circularly polarized patch antenna as the side channel. All X310 are synchronized with a shared external clock (GPSDO). The sampling rate is set to 2 M/s.

5.4 Synchronization of Message Queues

The server maintains four queues to hold the messages reported by the three gateways and the OPS. Fig. 7 sketches the message structures for the gateway and label messages.

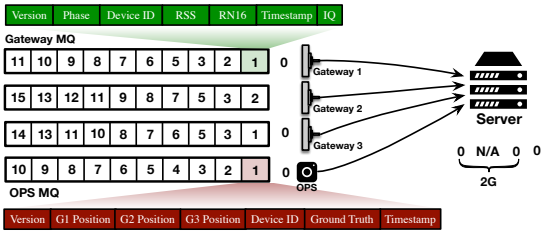


Fig. 7: Illustration for message queues and the message structure. The server maintains four message queues for the three gateways and the OPS, respectively. The queues are out of synchronization.

Specifically, the gateway message includes an array of `Phase` values and `RSS` values, the `DeviceID`, `Timestamp`, and the raw `IQ` data. The OPS message includes the `Positions` of the three gateways, `Timestamp`, and the `GroundTruth` of the device's location. When receiving a packet from the device of interest, the whole system samples its location. Ideally, the packet should be received by the three gateways concurrently, and thus the queues are rigorously synchronized on account of the timestamp. However, they are usually misaligned in practice for two reasons. First, our antenna arrays are equipped with directional elements with a ±45° beamwidth. One array might not receive the packets when the tag moves out of its beam, but the other arrays work correctly. Second, the OPS reports the device's location with a rate of 120 Hz, which is far higher than the reading rate of an RFID system. Thus, the queue of OPS receives more packets than gateway queues.

To address this issue, we firstly use the NTP to synchronize the timestamp among the devices. The redundant messages from the OPS queue are dropped if their timestamps are 90 ms earlier than the head messages of gateway queues. Finally, the grouped messages are saved into the database (e.g., MongoDB [38]). We will grant third-parties to download or connect to this database.

6 RAY DATABASE

With the help of the acquisition platform, we set up a cross-scene, large-scale, well-annotated, and 3D localization database, named *Ray*, which aims to serve as a benchmark dataset for deep learning on indoor localization. To this end, we collect data from 14 different types of scenes with 37 different settings. At present, the database contains 1.32 million location records, all of which are annotated with millimeter-accuracy labels. Each record is a sampling of the device's location and thereby contains the device ID, the position label, an array of phase values, RSS values, the gateway information, timestamp, and the raw IQ baseband signal. The summary of the collected data is shown in Table 2.

6.1 Data Scale

The Ray database contains 1,322,682 records (including location labels and other information). As aforementioned, the

Table 2: Summary of Ray Database

Env. (#)	Scene (#)	Setting (#)	RSS (dBm)	1G (#)	2G (#)	3G (#)	Total (#)	Density (p/m^3)	Space (m^2)	MP. (#)	Dist. (m)	Size (GB)	DR. (r/s)	TS. (min)	Temp. (°C)
Semi-Indoor Env.	A	**S1**	−62.5	32,191	35,308	16,893	84,392	3,843.0	78.5	10	5	48.24	40.4	79.5	31.2
		S2	−66.4	21,551	24,480	11,300	57,311	4,689.9	314.2	8	10	28.47	38.0	44.4	30.3
		S3	−66.7	20,372	23,773	11,382	55,527	5,274.2	706.9	7	15	26.88	42.4	41.2	29.9
		S4	−69.4	19,483	24,858	10,177	54,518	3,787.0	1,256.6	9	20	24.18	41.0	35.1	29.4
		S5	−71.0	16,567	22,278	11,457	50,302	4,336.4	1,963.5	10	25	25.38	40.1	35.0	27.2
		S6	−75.0	18,414	21,818	11,009	51,241	5,865.5	2,827.4	9	30	22.68	38.2	32.9	27.4
		S7	−77.4	16,445	22,807	12,037	51,289	5,871.0	3,848.5	10	35	23.82	35.9	33.8	27.7
		S8	−78.8	28,672	35,540	10,309	74,521	7,834.4	5,026.5	10	40	35.28	40.5	50.8	28.1
		S9	−79.3	28,672	35,540	10,309	61,909	4,235.7	6,361.7	8	45	27.66	41.4	40.9	28.3
		S10	−79.1	32,919	33,753	9,803	76,475	5,223.7	7,854.0	13	50	25.80	41.2	63.6	29.0
		S11	−88.6	29,440	17,387	3,359	50,186	10,490.4	7,854.0	11	55	18.15	31.5	29.8	28.7
	B	**S12**	−71.8	4,301	10,683	8,044	23,028	538.7	1,963.5	10	25	43.62	6.2	69.5	30.1
		S13	−76.9	6,245	10,172	4,940	21,357	702.1	3,848.5	13	35	39.07	5.9	66.9	30.4
		S14	−78.1	6,942	19,579	11,782	38,303	1,381.8	5,026.5	12	40	64.44	6.8	113.1	30.9
	C	**S15**	−68.3	5,533	9,075	4,118	18,726	6,079.9	1,256.6	16	20	40.02	7	74.5	33.1
	D	**S16**	−68.9	24,007	32,448	21,083	77,538	4,345.1	530.9	7	13	21.48	33.9	39.5	29.2
		S17	−67.0	15,684	16,584	8,303	40,571	2,545.9	530.9	10	13	46.92	32.2	81.6	28.8
Full-Indoor Evn.	E	**S18**	−66.2	2,326	31,693	126,475	160,494	38,212.9	314.2	14	10	89.93	20.7	61.2	18.4
		S19	−65.3	8,720	69,915	N/A	78,635	27,924.4	314.2	11	10	31.25	27.4	124.1	24.9
		S20	−63.7	3,173	26,930	N/A	30,103	10,906.9	314.2	11	10	12.19	16.1	26.5	25.1
		S21	−64.9	2,998	23,918	N/A	26,916	8,900.8	314.2	10	10	11.74	39.3	21.7	27.6
		S22	−65.4	18,872	13,170	N/A	32,042	5,057.1	314.2	11	10	12.47	40.7	18.0	24.8
	F	**S23**	−65.1	4,930	17,714	25,823	48,467	22,627.0	153.9	9	7	71.46	8.1	124.1	25.8
	G	**S24**	−61.4	1,749	4,222	4,550	10,521	4,911.8	78.5	8	5	11.88	4.2	65.9	27.5
		S25	−60.2	891	2,425	1,975	5,291	937.8	78.5	5	5	9.42	4.3	33.6	30.1
		S26	−61.9	1,911	3,973	839	6,723	2,394.2	78.5	12	5	11.94	5.3	31.0	27.3
		S27	−61.9	1,593	3,527	3,293	8,413	1,828.9	78.5	10	5	17.28	4.1	100.1	28.2
		S28	−63.7	1,297	3,526	4,088	8,911	1,600.4	78.5	6	5	13.92	9.6	57.5	27.9.
	H	**S29**	−61.7	1,026	1,984	2,092	5,102	912.7	113.1	7	6	14.11	3.7	33.1	29.3
		S30	−60.4	526	1,335	1,789	3,650	1,011.6	113.1	5	6	9.72	5.1	17.8	28.8
	I	**S31**	−60.9	964	2,555	3,947	7,466	823.0	113.1	7	6	16.02	8.9	28.2	29.9
		S32	−61.0	528	1,244	982	2,754	307.9	113.1	9	6	12.72	5.4	17.2	28.0
	J	**S33**	−61.6	9,379	16,164	N/A	25,543	1,576.7	78.5	8	5	8.41	48.3	11.8	18.5
	K	**S34**	−71.0	25,988	2,894	N/A	28,882	1,380.6	1256.6	4	20	11.78	29.6	16.6	17.8
	L	**S35**	−77.7	23,743	28,891	N/A	52,634	935.9	1963.5	10	25	23.85	27.8	34.4	17.6
	M	**S36**	−79.9	8,874	12,809	N/A	21,683	1,335.2	3217.0	8	32	8.53	24.0	12.3	17.3
	N	**S37**	−68.5	14,250	7,479	N/A	21,729	848.1	254.5	6	9	7.83	30.6	11.3	16.8

RSS is the signal strength; **1G**, **2G**, and **3G** are the labels indicating how many gateways can receive the packets concurrently at the corresponding locations; **Den.** is the density of locations in the unit of cubic meter; **Space** is the area of the surveillance region; **MP.** is the average number of multipath propagations; **Dist.** is the average distance from the center of the collected locations to the three gateways; **Size** is the volume of the raw baseband signal in the unit of GB; **DR.** is the data rate defined as the number of collected packets per second; **TS** is the time span that how many minutes were taken to collect these data; Temp. is the average temperature in the scene.

packets from the target device might not be received by the three gateways concurrently. In the collected data, we label the location with 1G, 2G, and 3G to indicate how many gateways can receive the packet at the location concurrently. For example, there exist 35,308 locations labeled with 2G in S1, which means the packets transmitted at the 25,308 locations can be received by two of the three gateways. In the Ray database, a total of 378,942, 604,214, and 352,158 locations are labeled with 1G, 2G, and 3G, respectively. They occupy 28.65%, 45.68%, and 26.62% of the records. Two directions are sufficient for triangulation to locate a device in 3D. Thus, 72.3% records are functional for the triangulation-based localization. The remaining 28.65% records can still be used for signature-based localization. In addition, the Ray database also provides roughly 15 TB raw IQ baseband signals received from RFID tags, which can be used for a benchmark dataset for the phase estimation algorithms or decoding algorithms. Particularly, the data collected from three 4×4 planar uniform arrays can be scaled down to train for any smaller-scale arrays (e.g., 2×2) via removing the superfluous element data.

6.2 Scene Diversity

As shown in Fig. 8, the data were collected in 2 main environments (full-indoor and semi-indoor environments.), 14 scenes (A∼N), and 37 settings (S1∼S37). The largest scene covers 7,854 m^2, and the smallest scene is roughly 78.5 m^2.

• **Semi-indoor Environment**. We first collect the data in a large-area semi-indoor environment, with the purpose of quantifying the impact of the distance. In such an environment, our acquisition platform is deployed in four scenes labeled A, B, C, and D, which are large-area and semi-closed

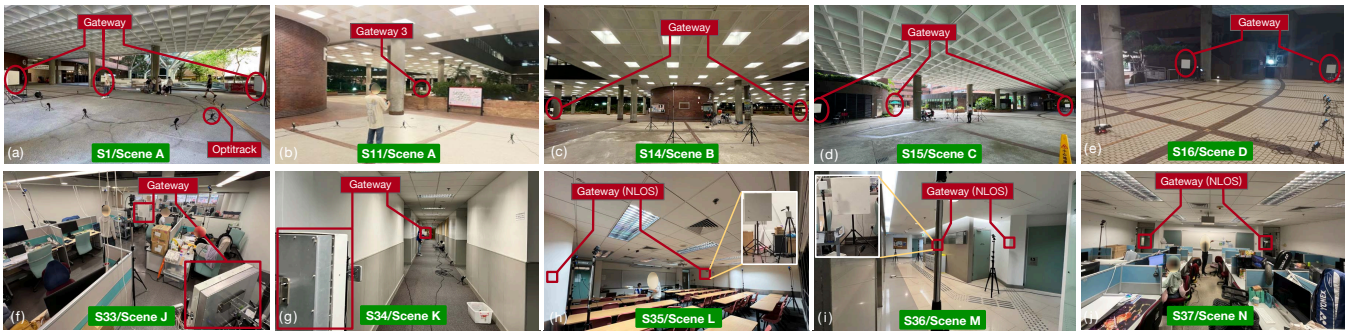
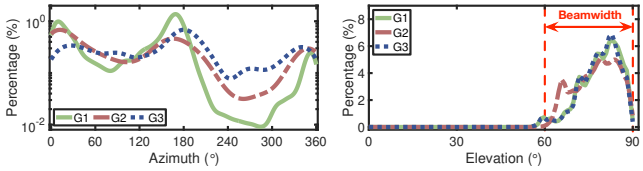


Fig. 8: Illustration of some example scenes. (a)-(e) shows the semi-indoor environment, which are large-sized and semi-closed halls. (f)-(j) show the full-indoor environment.



(a) Azimuth Distribution (b) Elevation Distribution

Fig. 9: Illustration of angle distribution for dataset #S29

halls, as shown in (a)-(e) of Fig. 8. In these scenes, 15 distance settings (i.e., S1-S15) were considered where the distance varies from 5 to 50 m. The distance is the mean value of that between the scene center to the three gateways. These data can serve to train the neural networks for the localization in large-sized scenes (e.g., shopping malls, stadiums, halls, playgrounds, and so on).

• **Full-indoor Environment**. We then collect the data in the full-indoor environment. We deploy the platform in 10 rooms (i.e., Scenes E-N) with 20 settings (i.e., S18-S37). Scene E is a warehouse, Scene F is a lab room, Scene G-I are classrooms. Scene J and N are offices, Scene K is the hallway, Scene L is a meeting room, and Scene M is a lift lobby, as shown in (f)-(j) of Fig. 8. The distance in these 20 settings ranges from 5 to 32 m. Particularly, the gateways are deployed behind the wall in Scenes L, M, and N. The data are especially suitable for study on localization in the NLoS scenario. Majority of these data are collected in the scenes full of people passing by and various metal reflectors.

In each scene, we might repeat to collect the data with different settings. A setting refers to the explicit or implicit influence from the gateway layout, distance, temperature, collection time, or other unknown factors. We believe these data collected in different settings is of great value for studying transfer learning (i.e., transfer a neural network learned in one setting to work for another one).

6.3 Spatial Distribution

The learning performance highly depends on the quality of labels, especially the sample distributions. Next, we present an analysis on the label distributions.

• **Directional Distribution**: Triangulation is based on the directions. Thus, the distribution of the sampled directions significantly impacts the learning performance. We choose the S29 dataset collected in Scene H (a classroom) as an example to show the azimuth and elevation angle distribution of location labels. The results are shown in Fig. 9. The two angles are defined on the basis of the LCS as shown Fig. 1. Fig. 9(a) shows the azimuth distribution where most azimuth angles are visited uniformly except for the ones ranging between 240 $\sim$ 300°. In this range, the RFID tag is close to the floor and hardly powered up due to the Fresnel zone effect (i.e., RF signals reflected off the floor create strong self-interference), leading to a relatively few percent of samples (i.e., 0.1 $\sim$ 0.2%). Fig. 9(b) shows the elevation distribution where the samples concentrate on the range of 60° $\sim$ 90° because our array elements are directional antennas with ±30° beamwidth, out of which RF signals cannot be received.

• **Coverage**: We adopt the OptiTrack comprising 12 PrimeX infrared cameras [34] as the OPS to collect the location labels. These cameras can cover approximately $6 \times 6 \times 2\ m^3$ space. Constrained to this range limit, the gateway must be moved gradually away from the surveillance region in the semi-indoor scenes to collect the long-range localization labels. Specifically, in the Scene A, the cameras are deployed in a circle with a radius of 2.5 m, inside which the device moves freely. Meanwhile, the three gateways are pulled away from the center of the circle with a step of 5 m in the 11 settings (S1-S11). In this way, we can cover the whole $7,854\ m^2$ maximum area. To show a big picture about the coverage, we transform the locations collected in the 11 settings labeled with the individual global coordinates to a common polar coordinate system (PCS). The results are shown in Fig. 10. It can be seen that the samples are well distributed in the surveillance area with a very satisfying full coverage.

6.4 Scene Complexity

We conduct a simple analysis of the scene complexity, which is reflected from the average number of multipath in a scene.

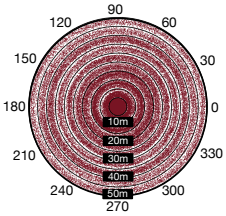


Fig. 10: Labels on PCS

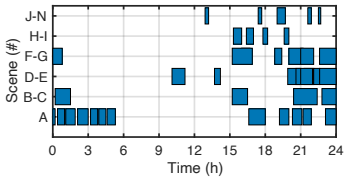


Fig. 11: Acquisition time

To this end, we firstly use the MUSIC [31] algorithm to generate the spatial spectrums and then count the number of local peaks. Ideally, the whole spatial spectrum should only spike at a single direction pointing to the LoS path. However, the spectrum contains many local spikes, each viewed as a potential NLoS path. We use the number of local peaks to indirectly reflect the scene complexity. Note that the direct number of the multipath propagation is unclear because two nearby local peaks might fuse into one. We randomly select 10,000 locations from each dataset, and report the average number of their complexities. The results are shown in Table 2. Surprisedly, the semi-indoor scenes are more complex than the full-indoor scenes. As Fig. 8 shows, the humpy and hillocky ceilings of the semi-closed halls entangle the propagations even if they lack walls. In short, the semi-indoor complexity varies from 8 to 13, while that of full-indoor complexity changes from 5 to 15.

6.5 Physical Signals

In the database, we also store the raw I/Q baseband signal. Table. 2 shows the statistics on the average RSS of the backscatter signals from the 37 datasets. The received signals from a backscatter system are usually far lower than an active system because the nodes are battery-free. The highest RSS is about -50 dBm, and the lowest value is near to -90 dBm (i.e., limited to -90 dBm sensitivity of USRP). The range of a commercial RFID system is roughly 15 m around. However, we extend this range to 50 m using a bistatic mode where the RF source and the receiver (i.e., antenna arrays) are separated, and the former is deployed 2 m away from the tag to provide sufficient power.

6.6 Miscellaneous

Finally, we show the acquisition time based on the record timestamps in Fig. 11. The data are mainly collected from 15:00 to 21:00 in our campus when many students go across the scenes. The longest acquisition took almost 2.06 hours in S19 and S23. Such information might be helpful to study the temporal impact on the performance. In addition, the average environment temperature during the acquisition is shown in Table 2. Many previous works [39–41] claimed that the phase value of RF signals received from an RFID tag clearly varies as a function of the temperature. Since the localization is based on the phase, the performance of neural network might be affected by the temperature. We collected the data with the minimum temperature of 16.8 °C and the maximum of 33.1 °C. Such temperature information can serve the related work.

7 ThreeBodyNet: A DEEP NEURAL NETWORK FOR LOCALIZATION

In this section, we propose a deep neural network, called *ThreeBodyNet*, to verify the effectiveness of the database and to demonstrate how the database can be used.

7.1 Preprocessing

We conduct two important preprocessing methods over the raw RF signals before feeding them into the neural network.

•**Smoothing Phase Sequence**. When acquiring the backscatter signals using the gateways, the target tag was continuously moved in the space. The received phase might receive the negative influence from the movement (i.e., Doppler effect) or other unknown instantaneous interference. To defend against the severe fluctuation in phase, we firstly use the Kalman filter to smooth the phase values acquired by each antenna element. The motion is assumed to be uniform and thereby the phase should change uniformly. Therefore, we adopt the similar Wiener velocity model or continuous white noise acceleration (CWNA) model as used in [42] to set up the Kalman filter.

•**Generating Spatial Spectrum (SS)**. Then, the server generates the spatial spectrums on the basis of Eqn. 2 using an array of phase, including 16 smoothed values. The angular resolution of spatial spectrum is set to 1° by default. The core of our neural network is based on the CNN. Thus, we convert the spatial spectrum into 360×90 pixels where the X-axis and Y-axis represent the azimuthal and elevation angles, respectively. Particularly, the azimuthal angle is repeated every 360°. Projecting SS into an image losses the azimuthal continuity in the left and right margins when conducting the convolution. Thus, we attach a cyclic prefix and a cyclic suffix on the left and the right margins, respectively. They are the copies of the last or the first columns of pixels. As a result, the spectrum image becomes 380×90 pixels in size. In addition, we use red (R), green (G), and blue (B) to color the spatial spectrums generated from Gateway 1, 2, and 3.

Mathematically, the spatial spectrum is the measured Ω of Eqn. 10. We might have one, two, or three Ω as input to feed the following neural network, and wish it to find the optimized resolution, i.e., the position of the device.

7.2 Network Architecture

The two estimated directions are enough to locate a target in a 3D space by using triangulation. It seems that our three gateways can provide a redundant measurement. However,

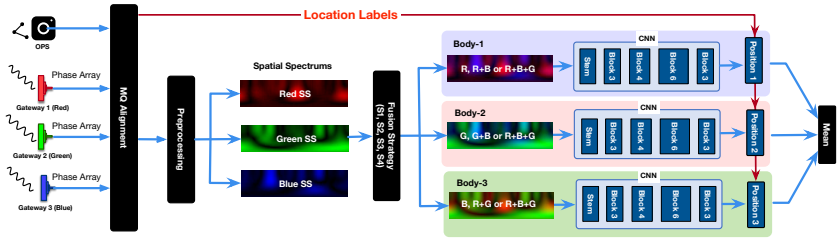


Fig. 12: Architecture of ThreeBodyNet. The ThreeBodyNet is used to fuse the measurements taken by the three gateways. It contains three subnetworks (called bodies), each of which is a ResNet-50 based convolutional network. For clarity, we use red, green and blue to color the spatial spectrums from the three gateways, respectively. Each body accepts a single-channel SS or a multi-channel SS as input and outputs the target position.

Table 3: Configuration of the body network

Name	Layer	Kernel Size	Output	Param.
Input	SS	N/A	(3,90,360)	0
Stem	1 Conv.	(3,3,64)	(64,47,182)	1.8k
	MaxPools	(3,3)	(64,47,91)	0
Block 3	4 Conv.	(1,1,64), (3x3,64) (1,1,64), (1,1,64)	(47,91,256)	74.8k
	(3 Conv.) × 2	(1,1,256), (3,3,64) (1,1,64)	(47,91,256)	140.8k
Block 4	4 Conv.	(1,1,256), (3,3,128) (1,1,128), (1,1,256)	(24,46,512)	379.4k
	(3 Conv.) × 3	(1,1,512), (3,3,128) (1,1,128)	(24,46,512)	840.2k
Block 6	4 Conv.	(1,1,512), (3,3,256) (1,1,256), (1,1,512)	(12,23,1024)	1.5m
	(3 Conv.) × 5	(1,1,1024), (3,3,256) (1,1,256)	(12,23,1024)	5.6m
Block 3	4 Conv.	(1,1,1024), (3,3,512) (1,1,512), (1,1,1024)	(6,12,2048)	6.0m
	(3 Conv.) × 2	(1,1,2048), (3,3,512) (1,1,512)	(6,12,2048)	8.9m
Output	AvgPools+FC	(6,12)	3	6.1k

the common area covered by the three gateways is highly limited, such that only 26.62% records are labeled 3G. Even if 45.68% records are labeled with 2G, they might be collected by any two of the three gateways, resulting in three combinations (i.e., R+G, R+G, B+G). Moreover, each spatial spectrum is created in an LCS, but the ground truth in a common GCS is labeled. Thus, these spatial spectrums cannot be simply mixed together to train a single neural network.

To address the challenges and maximize the utilization of the database, we propose a DNN, called *ThreeBodyNet*, which comprises three bodies, each of which is a subnetwork, as shown Fig. 12. Each body is constructed by using a ResNet-50 convolutional neural network. The three bodies are used to fuse the spectrum images generated from the three gateways. They have the exact same structures but different network parameters. The detailed configurations of the body network are listed in Table 3. For each convolutional layer in the block, we add a batch normalization layer, and use Rectified Linear Unit (ReLU) as the activation function. Each body network is trained individually using the Mean-Squared loss function of $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(p_i - \hat{p}_i)^2$ where p_i is the predicted position that the body network output and the label, $\hat{p}_i$ is the location label (ground truth), and N is the batch size. The three-body networks independently predicate three position results. If multiple results are predicted, the mean is reported.

7.3 Training Strategy

We can train and use the ThreeBodyNet with four potential strategies. *Strategy 1*: the three bodies accept the SSs from the red, green, and blue gateways, respectively. Each body uses a single SS to output the position. In this way, all records, including 1G labels, can be used. The device can be located once at least one gateway receives its packet. Essentially, this strategy uses the signature-based localization because one SS cannot take the triangulation in theory. *Strategy 2*: this strategy only uses the 2G records. As two color channels, the two corresponding SSs are combined into a single image. In this way, three possible results can be obtained (i.e., R+B, R+G, and G+B), each of which is fed into an individual body network. Consequently, once at least two gateways receive the packet, the device can be located. The body network utilizes two SSs acquired by different gateways, so the strategy is indeed for the desired triangulation. *Strategy 3*: this strategy only uses the 3G records. As three color channels, the three corresponding SSs are combined into a single image. These images are repeated to be fed into three bodies for training. The two redundant bodies are used to reduce the uncertainty caused by the learning randomness. The device can be located only when its packet must be acquired by three gateways concurrently. *Strategy 4*: this strategy is similar to the Strategy 2 but only uses 3G records, which is split into three combinations (R+B, R+G and B+G). They are further merged into a single image as the input of the following body. Unlike Strategy 2, three predicated results are output by the whole network each time.

8 EVALUATION

The ThreeBodyNet is developed using the Pytorch framework. We use an 80/20 training/test split on Ray database. Our model is trained on a machine with an AMD 5900x (4.9GHz) processor, 64 GB RAM, and an NVIDIA 3080Ti GPU. We use the Stochastic Gradient Descent (SGD) optimizer with a learning rate of e^{-4} and momentum of 0.9. We maintain a batch size of 64 when training our model. It takes roughly 10 hours to train 1,200 epochs. However, it only takes 0.1 seconds for each test. In this section, we will report the performance of the ThreeBodyNet.

8.1 Accuracy of Training Strategy

First, we would like to know the advantage and disadvantage of four training strategies. To this end, we respectively train the ThreeBodyNet using the four strategies across the S1 dataset. The CDFs of accuracy are shown in Fig. 13. The median errors are 14.99 cm, 12.43 cm, 8.52 cm, and 6.86 cm by using Strategy 1, 2, 3, and 4. Compared with the Strategy 1, Strategies 2, 3, and 4 can improve the accuracy by 17.1%, 43.2%, and 54.2%, respectively. The results are not surprising because the latter three strategies provide higher accuracy at the cost of more measurements. The accuracy of Strategy 1 is unsatisfying in that it is based on the signal signature, i.e., a location is supposedly related to a unique spatial spectrum.

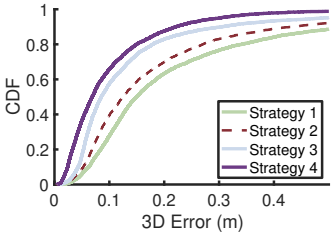


Fig. 13: Accuracy vs. Strategies

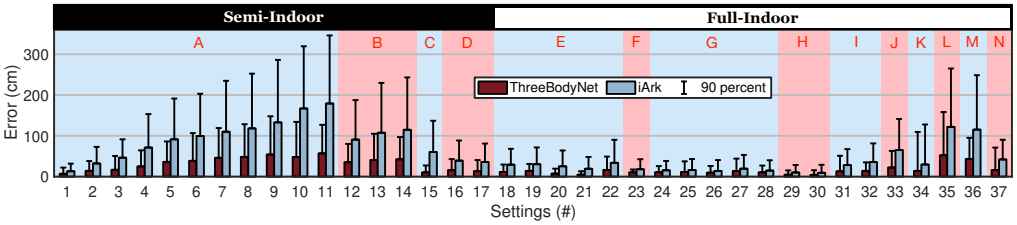


Fig. 14: Accuracy vs. Settings

Although Strategies 3 and 4 perform best, they can only work within a limited shared surveillance region that must be covered by three gateways concurrently. Specifically, Strategy 1, 2, 3 and 4 are applicable to the area of 100%, 61.9%, 20.1% and 20.1%. Thus, the Strategy 2 is highly recommended in practice as a trade-off between the coverage and the accuracy. Other datasets also show the similar results. Thus, Strategy 2 is used by default in the following experiments.

8.2 Accuracy of 3D Localization

We train ThreeBodyNet using the 37 datasets (i.e., S1-S37) accordingly and report their accuracy results in Fig. 14. The training totally took about 400 hours. From the figure, we have the following findings.

■ **Accuracy vs. Distance**. In the settings of S1-S11, the distance is increased from 5 m to 55 m, and the corresponding RSS degrades from -62.6 dBm to -88.6 dBm. The nearly 29 dB attenuation forces median error to increase from 6.8 cm to 54.3 cm, with a near 7× increase. Namely, the error is increased by 1.8 cm per 1 dB RSS attenuation. In essence, the impact of the distance can be considered as that of the RSS since the input power is fixed in our scenario. The impact of RSS is also reflected from full-indoor datasets (S18-S37), in which the mean RSS is with -65.1 dBm an std of 3.7 dBm. As a result, the median error in the full-door environment is maintained at 12.2 cm with a 90th percentile error of 33.1 cm. Clearly, the RSS (or equivalent distance) plays a key role in the localization, especially in a backscatter system where the RSS of a battery-free backscatter is far lower than that of an active device. It is more difficult to localize a backscatter device than an active one in practice.

■ **Semi-Indoor vs. Full-Indoor**. For fairness, we only consider the datasets of S1-S3 and S15-S17 collected in the semi-indoor environments, and the dataset of S18-S34 collected in the full-indoor environments. In these datasets, the distance is less than 15 m, and the RSS is at a similar level (i.e.,-70 dBm above). We use them for the accuracy comparison. As desired, ThreeBodyNet shows the comparable performance in these datasets with a median accuracy of 10 cm. This demonstrates that ThreeBodyNet can serve as a general-purpose learning network working for different environments full of various multipath propagations.

■ **Blockage**. The datasets of S35-S37 are collected in blockage scene where the gateways are deployed behind walls and 25 m, 32 m, and 9 m away from the device. The RSS is reduced to -77.7 dBm, -79.8 dBm, and -68.5 dBm. The condition is comparable to S5, in which the distance is also 25 m, and the RSS is -71 dBm. Clearly, the wall increases an addition roughly 6 dB attenuation, which is also confirmed in [43]. Consequently, the median errors with in S35 and S5 are 52.3 cm and 36.02 cm, respectively. Clearly, the blockage seriously affects the accuracy. It might double the error.

■ **Comparison with State of the Art**. We also compare the performance of the ThreeBodyNet with iArk [21]. iArk is a similar platform that uses a DNN and an antenna array for triangulation. We run iArk's neural network across our Ray database. The results are shown in Fig. 14. Notably, ThreeBodyNet always outperforms than iArk in all settings. Specifically, the accuracy of the ThreeBodyNet is improved by 56.4% on average compared with the iArk. The outperformance is mainly attributed to the separated three body networks and multiple gateways.

8.3 Impact Analysis

Next, we would like to evaluate the accuracy with the impacts from various parameters.

■ **Impact of Array Size.** Apart from the whole 4×4 array, we also evaluate the accuracy using 3×3 and 2×2 arrays. The smaller arrays are created by ignoring the data collected via the needless elements. The accuracy results are plotted in Fig. 15. The median errors are 14.62 cm, 13.15 cm, and 11.44 cm, and the 90th percentile errors are 25.98 cm, 33.27 cm, and 40.03 cm, respectively. The results are in line with the theory that large-sized arrays provide a larger aperture and thus gain higher accuracy.

■ **Impact of Preprocessing**. Next, we evaluate the accuracy regarding the proposed phase estimation algorithm and the two preprocessing methods. In the experiment, the S4 dataset is chosen to train the ThreeBodyNet using four types of data, PE, LNPE, LNPE+KF, and LNPE+KF+CC. The PE and LNPE are the two phase estimators. PE is proposed in [44], while LNPE is the low noise phase estimator proposed in this work. The median localization errors using PE and LNPE are 63.3 cm and 38.8 cm, respectively, where the 38.7% improvement mainly attributes to the dual-channel design that can

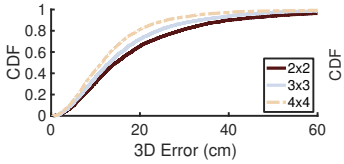


Fig. 15: Accuracy vs. Array Size

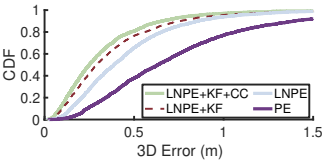


Fig. 16: Accuracy vs. Preprocessing

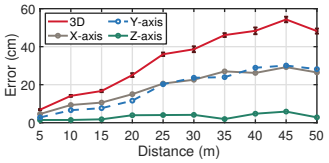


Fig. 17: Accuracy vs. Axis

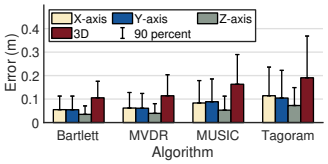


Fig. 18: Accuracy vs. Spectrum

help counteract many unknown negative influences. Based on the LNPE, we also consider the effect of the two preprocessing methods. The median error of LNPE+KE is 29.0 cm, which gives another 25.2% improvement by excluding the instantaneous interference using Kalman Filter (KF). Finally, we apply the cyclic prefix and cyclic suffix (CC) in the spatial spectrum. As a result, the median error of LNPE+KE+CC is further reduced to 25.1 cm with a 13% improvement. CC helps only when the direction of the device is located near to the margins of the spatial spectrum.

■ **Impact of Axis**. We also compare the errors at the X, Y, and Z axes regarding the datasets of S1-S13. The results are shown in Fig. 17. The errors at the X and Y axes are comparable, which are increased from 4.6 cm and 2.8 cm to 29.2 cm and 30.1 cm, respectively. The error at Z-axis (from 1.42 cm to 5.84 cm) seems far smaller than that at the other two axes. This is because the device was moved along the Z-axis within a 2 m small range, limited to the hall height.

■ **Impact of Spectrum Algorithm**. Since the input of the DNN are images converted from spatial spectrums, we evaluate the impact of different spectra generation algorithms on localization accuracy. We select four classical algorithms to test our DNN: Bartlett [29], MVDR [30], MUSIC [31], and Tagoram [1]. Bartlett is the most primitive algorithm mentioned early. MVDR adaptively maximizes the SINR in the desired direction and suppresses the signal from other directions. MUSIC decomposes the received signal into two orthogonal subspace and estimates the direction. Tagoram proposed DAH with a Gaussian kernel to eliminate the phase offset. By using a triangulation approach, Tagoram and MUSIC algorithms achieved a higher accuracy than the others. The S23 dataset is selected to train the ThreeBodyNet. Fig. 18 plots the 50th and 90th percentile location errors at X, Y, Z axis and a combined 3D error. The median 3D errors of four algorithms are 10.5 cm, 11.4 cm, 16.3 cm and 19.1 cm; 90th percentile errors are 17.6 cm, 20.4 cm, 29.0 cm and 36.8 cm, respectively. Surprisedly, the most naive algorithm can achieve a highest accuracy by using ThreeBodyNet. We guess that this naive spectrum possibly reserves more original information and benefits the neural network to find the pattern.

9 RELATED WORK

Our work falls under broad indoor localization studies.

Indoor RF Localization. RF localization is a long-studied topic with extensive works [1, 45]. Various metrics of RF signals are widely used for localization: RSSI [45], carrier phase [1, 46], channel state information (CSI) [3, 47], Time-of-Flight (ToF) [48], and AoA [4, 21, 25, 49–52]. Here we focus on AoA based methods, which have been widely used for various scenarios, including Wi-Fi [4, 25, 50–52], Bluetooth [49, 53], RFID [21, 54, 55], and LoRa [56]. AoA-based localization algorithms suffer from problems caused by the environment (such as complex multipath effect, non-line of sight, and so on.). These problems have been extensively studied in literature [18, 25]. Past works solved this problem by identifying the direct path and trying to eliminate the effects caused by the environment [3, 49]. mD-Track [3] improves the AoA accuracy under multipath scenarios by using ToF measurement for direct path identification. However, this approach fails when the multipath signals are extremely close or the direct path is completely blocked. ThreeBodyNet uses deep learning to comprehensively understand the environment so that to locate even under NLoS case.

Deep learning for localization. Unlike model-based algorithms, recent state-of-the-art works focused on training a deep learning model to improve the accuracy in complex indoor scenarios [21, 22, 25]. Meanwhile, many localization datasets are proposed for the need of learning models [20–22, 25]. We differ from past works in two aspects. First, we propose the largest dataset with high accuracy labels and diverse scenarios. This allows us to train and evaluate the localization model on an unprecedented scale. Second, we propose a novel deep learning-based distributed signal processing architecture. Most past works only applied the deep learning model for enhancing single station AoA estimation and combining distributed measurements with the geometry method [25]. We break this stereotype and use a unified model. ThreeBodyNet improves the accuracy and also saves the coordination calibration.

10 CONCLUSION

We introduce an unprecedented ultra large-sized localization database, Ray, for indoor localization. The database contains over one million positions with millimeter-accuracy labels. To this end, we build a customized hardware platform, developed sophisticated baseband processing algorithms, and propose the ThreeBodyNet– a localization-oriented DNN. We believe that the Ray's rich structure and dense coverage may help further understanding of the wireless environment and the large-scale indoor localization with the DNN.

REFERENCES

[1] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-time tracking of mobile rfid tags to high precision using cots devices," in *Proc. of ACM MobiCom*, 2014, pp. 237–248.

[2] Y. Ma, N. Selby, and F. Adib, "Minding the billions: Ultra-wideband localization for deployed rfid tags," in *Proc. of ACM MobiCom*, 2017.

[3] Y. Xie, J. Xiong, M. Li, and K. Jamieson, "md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–16.

[4] Y. Xie, Y. Zhang, J. C. Liando, and M. Li, "Swan: Stitched wi-fi antennas," in *Proc. of ACM MobiCom*, 2018.

[5] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3d tracking via body radio reflections," in *Proc. of USENIX NSDI*, vol. 14, 2013.

[6] F. Adib, Z. Kabelac, and D. Katabi, "Multi-person motion tracking via rf body reflections," in *Proc. of USENIX NSDI*, 2015.

[7] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "Rf-based 3d skeletons," in *Proc. of ACM SIGCOMM*, 2018, pp. 267–281.

[8] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proc. of IEEE CVPR*, 2018, pp. 7356–7365.

[9] Y. Ma and E. C. Kan, "Accurate indoor ranging by broadband harmonic generation in passive nltl backscatter tags," *IEEE Transactions on Microwave Theory and Techniques*, vol. 62, no. 5, pp. 1249–1261, 2014.

[10] X. Hui and E. C. Kan, "Radio ranging with ultrahigh resolution using a harmonic radio-frequency identification system," *Nature Electronics*, vol. 2, no. 3, p. 125, 2019.

[11] A. Haniz, G. K. Tran, K. Saito, K. Sakaguchi, J.-i. Takada, D. Hayashi, T. Yamaguchi, and S. Arata, "A novel phase-difference fingerprinting technique for localization of unknown emitters," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8445–8457, 2017.

[12] M. Youssef and A. Agrawala, "The horus wlan location determination system," in *Proc. of ACM MobiSys*, 2005, pp. 205–218.

[13] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are facing the mona lisa: Spot localization using phy layer information," in *Proc. of ACM MobiSys*, 2012, pp. 183–196.

[14] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proc. of ACM MobiCom*, 2012, pp. 269–280.

[15] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Push the limit of wifi based localization for smartphones," in *Proc. of ACM MobiCom*, 2012, pp. 305–316.

[16] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: Unsupervised indoor localization," in *Proc. of ACM MobiSys*, 2012, pp. 197–210.

[17] L. Ni, Y. Liu, Y. Lau, and A. Patil, "Landmarc: Indoor location sensing using active rfid," *Wireless networks*, 2004.

[18] J. Wang and D. Katabi, "Dude, where's my card?: Rfid positioning that works with multipath and non-line of sight," in *Proc. of ACM SIGCOMM*, 2013.

[19] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for wifi-based indoor localization," in *Proc. of ACM AAAI workshop*, vol. 6, 2008.

[20] U. Raza, A. Khan, R. Kou, T. Farnham, T. Premalal, A. Stanoev, and W. Thompson, "Dataset: Indoor localization with narrow-band, ultra-wideband, and motion capture systems," in *Proceedings of the 2nd Workshop on Data Acquisition to Analysis*, 2019, pp. 34–36.

[21] Z. An, Q. Lin, P. Li, and L. Yang, "General-purpose deep tracking platform across protocols for the internet of things," in *Proc. of ACM MobiSys*, 2020, pp. 94–106.

[22] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.

[23] N. Dvorecki, O. Bar-Shalom, L. Banin, and Y. Amizur, "Intel open wi-fi rtt dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/h5c2-5439

[24] Y. Shu, Q. Xu, J. Liu, R. R. Choudhury, N. Trigoni, and V. Bahl, "Indoor location competition 2.0 dataset," January 2021. [Online]. Available: https://www.microsoft.com/en-us/research/publication/indoor-location-competition-2-0-dataset/

[25] R. Ayyalasomayajula, A. Arun, C. Wu, S. Sharma, A. R. Sethi, D. Vasisht, and D. Bharadia, "Deep learning based wireless localization for indoor navigation," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[26] C. Li, Z. Cao, and Y. Liu, "Deep ai enabled ubiquitous wireless sensing: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.

[27] W. Qian, F. Lauri, and F. Gechter, "Supervised and semi-supervised deep probabilistic models for indoor positioning problems," *Neurocomputing*, vol. 435, pp. 228–238, 2021.

[28] C. Zhan, M. Ghaderibaneh, P. Sahu, and H. Gupta, "Deepmtl: Deep learning based multiple transmitter localization," in *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2021, pp. 41–50.

[29] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE signal processing magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[30] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *Signal Processing*, vol. 93, no. 12, pp. 3264–3277, 2013.

[31] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[32] "How AoA & AoD Changed the Direction of Bluetooth Location Services," https://www.bluetooth.com/blog/new-aoa-aod-bluetooth-capabilities/.

[33] T. Instruments, "The fundamentals of millimeter wave radar sensors."

[34] "OptiTrack," https://optitrack.com/, 2020.

[35] "RabbitMQ," https://www.rabbitmq.com/.

[36] "RF Switch," https://www.infineon.com/cms/cn/product/rf-wireless-control/rf-switches-spxt-dpxt/bgs18ga14/, 2016.

[37] "USRP X310," https://www.ettus.com/all-products/x310-kit/, 2020.

[38] "MongoDB," https://www.mongodb.com/.

[39] X. Wang, J. Zhang, Z. Yu, S. Mao, S. C. Periaswamy, and J. Patton, "On remote temperature sensing using commercial uhf rfid tags," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 715–10 727, 2019.

[40] S. Pradhan and L. Qiu, "Rtsense: passive rfid based temperature sensing," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 42–55.

[41] X. Wang, J. Zhang, Z. Yu, E. Mao, S. C. Periaswamy, and J. Patton, "Rfthermometer: A temperature estimation system with commercial uhf rfid tags," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[42] S. Sarkka, V. V. Viikari, M. Huusko, and K. Jaakkola, "Phase-based uhf rfid tracking with nonlinear kalman filtering and smoothing," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 904–910, 2011.

[43] E. C. Jones and C. A. Chung, *RFID and Auto-ID in Planning and Logistics: A Practical Guide for Military UID Applications*. CRC Press, 2016.

[44] R. Miesen, A. Parr, J. Schleu, and M. Vossiek, "360 carrier phase measurement for uhf rfid local positioning," in *2013 IEEE International Conference on RFID-Technologies and Applications (RFID-TA)*. IEEE,

2013, pp. 1–6.

[45] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "Landmarc: Indoor location sensing using active rfid," in *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003)*. IEEE, 2003, pp. 407–415.

[46] Y. Ma, N. Selby, and F. Adib, "Drone relays for battery-free networks," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, 2017, pp. 335–347.

[47] Z. Yang, Z. Zhou, and Y. Liu, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–32, 2013.

[48] A. T. Mariakakis, S. Sen, J. Lee, and K.-H. Kim, "Sail: Single access point-based indoor localization," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014, pp. 315–328.

[49] R. Ayyalasomayajula, D. Vasisht, and D. Bharadia, "Bloc: Csi-based accurate localization for ble tags," in *Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologies*, 2018, pp. 126–138.

[50] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *ACM SIGCOMM computer communication review*, vol. 45, no. 4. ACM, 2015, pp. 269–282.

[51] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system," in *Proc. of USENIX NSDI*, 2013, pp. 71–84.

[52] J. Gjengset, J. Xiong, G. McPhillips, and K. Jamieson, "Phaser: Enabling phased array signal processing on commodity wifi access points," in *Proc. of ACM MobiCom*, 2014, pp. 153–164.

[53] M. Cominelli, P. Patras, and F. Gringoli, "Dead on arrival: An empirical study of the bluetooth 5.1 positioning system," in *Proceedings of the 13th international workshop on wireless network testbeds, experimental evaluation & characterization*, 2019, pp. 13–20.

[54] J. Wang, D. Vasisht, and D. Katabi, "Rf-idraw: Virtual touch screen in the air using rf signals," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 235–246, 2014.

[55] J. Wang, J. Xiong, H. Jiang, X. Chen, and D. Fang, "D-watch: Embracing "bad" multipaths for device-free localization with cots rfid devices," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3559–3572, 2017.

[56] N. BniLam, D. Joosens, M. Aernouts, J. Steckel, and M. Weyn, "Loray: Aoa estimation system for long range communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2005–2018, 2020.